Information Bottleneck Problem Revisited

Farhang Bayat, Shuangqing Wei

Abstract—**In this paper, we revisit the information bottleneck problem whose formulation and solution are of great importance in both information theory and statistical learning applications. We go into details as to why the problem was first introduced and how the algorithm proposed using Lagrangian method to solve such problems fell short of an exact solution. We then revisit the limitations of such Lagrangian methods, and propose to adopt a more systematic method, namely, Alternate Direction Method of Multipliers (ADMM) to develop a more efficient ADMM algorithm with randomized permutation orders to solve such problems. More importantly, we mathematically demonstrate how our suggested method outperforms the original Information Bottleneck (IB) method. At the end, we provide numerical results to demonstrate the notable advantages our algorithm attains as compared with the well-known IB approach in terms of both attained objective function values and the resulting constraints. We further inspect the concepts of accuracy and convergence and the trade-off between them in our method.**

Index Terms—**information bottleneck, constrained optimization, convergence, accuracy, ADMM**

I. INTRODUCTION

The information bottleneck problem was first introduced as a trade-off problem between an input and an output through a median variable [1]. Namely, given two random variables U and X following a given joint distribution $P(U, X)$, we seek a mapping between X and Y to form a Markov Chain $U \rightarrow X \rightarrow Y$ under which the mutual information between X and Y is minimized under a constraint that the mutual information between U and Y is higher than a lower bound. A straightforward instance for such a model could be exemplified when we are observing X at one point, which shall be quantized and shared with another entity, who is interested in inferring about U using such quantized information of X. Thus, the cost due to distortion induced by compression is in terms of mutual information $I(U; Y)$. In this scenario, we hope that minimal rate is used to describe X through output Y, however we also want to make sure that an enough level of information about input U could be revealed through output Y.

In the information bottleneck method as detailed by [1], the above constrained optimization was posed as a non-convex problem with remarks of lacking a guaranteed globally optimal solution. The problem was then rewritten in the format of a Lagrange multiplier problem where the goal would be to minimize a linear combination of the two functions ($I(X; Y)$ and $I(U; Y)$) connected to one another through a Lagrange multiplier.

F. Bayat and S. Wei are with the School of EECS, Louisiana State University, Baton Rouge, LA 70803, USA (Email: swei@lsu.edu; fbayat1@lsu.edu). This work was supported in part by the National Science Foundation under Grant No. 1320351.

Such interpretation of the problem gave way to the use of convenient Lagrange multiplier methods and an iterative algorithm with every step of the algorithm specified in a simple formula. It was shown in [1] that the solution might not be optimal and instead they emphasized on the importance of the simplicity of the algorithm. This simplicity has resulted in the Information Bottleneck (IB) algorithm being widely used in many settings including both learning driven and information theoretic problems. However, missing was a more in-depth analysis of the original problem, the limitations of the intuitive algorithm proposed in [1] and the sacrifices it makes in order to make certain the algorithm is simple enough.

In this paper, we revisit the original information bottleneck problem in a constrained optimization setting. We go into details as to what the concerns of using the simplified Lagrange method are and not only offer a new algorithm of solving the problem to address such issues, but also (by going into details about the potential issues in the canonical IB approach) further justify the adoption of free and auxiliary variables as described in our new algorithm.

We then propose to adopt a more systematic method in optimization literature, namely Alternate Direction Method of Multipliers (ADMM) to develop a more efficient algorithm to solve such problems. This method offers the option of adapting penalty functions as a means of controlling the constraints imposed on the problem [2]. We showcase how such a method is superior to previous algorithms by both discussing the nature of two solutions and offering sufficient numerical evidence.

A. Related Works

Due to the nature of our paper, which heavily relies on discussing every step of the information bottleneck method, our main point of reference will be the original Information Bottleneck paper [1]. Still to showcase either the limitations of the IB method or the advantages of our suggested algorithm, we will refer to many other works throughout this paper.

We find it important to note that [1] was inspired by the works of [3] and [4]. Thus, from time to time, in order to indicate the inadequacies of [1]'s results, we will refer to these papers. Further comments on the inadequacies of [1] have been made in many works as recent as [5] which reflect the concerns of many previous authors who while aware of the limitations of IB, still chose to utilize it due to its simple implementation.

The idea of solving Lagrange multiplier problems where the goal is to either maximize or minimize a function using penalty functions has been explored in many previous mathematical and computing works such as [2]. These issues could be modeled in the form of an Augmented

Lagrange Multiplier (ALM) problem. However, due to the still-complex nature of solving any non-convex Lagrange multiplier problem (which requires carrying out gradient descent and checking for conditions on every Lagrange multiplier), further studies and methods are required. One algorithm of dealing with the limitations of an ALM problem is to instead try and adopt an ADMM algorithm. One of the most recent works about this method is [6] which details the superiority of adopting ADMM over continuing with ALM. [6] makes further observations on the accuracy of a non-convex Augmented Lagrange Multiplier problem which we will utilize to justify the numerical results achieved through this paper.

Finally, [7] offers a series of sufficient conditions on the desired utility function by which a convergence of the ADMM method could be guaranteed. We will offer insight into these sets of sufficient conditions and whether or not they are applicable to our problem settings.

B. Paper Contributions

In this paper, we (1) revisit the Information Bottleneck problem and the algorithm to solving it as suggested by [1]; (2) explain exactly if and how any step in the original algorithm could be problematic; (3) introduce the concept of augmented Lagrange multiplier problems to the same original problem; (4) showcase every possible superiority offered by the new method of looking at the problem; (5) offer an algorithm and an in-depth look at how it could be implemented to our problem; (6) demonstrate the practical results of running our suggested algorithm (ADMM) over the same problem; and (7) offer in-depth reasoning for the numerical results to further justify the novelty and importance of our new suggestion.

C. Paper Organization

The rest of this paper is organized as follows. In Section II, we introduce the system model and the original problem formulation. In Section III, we go into details as to how the original Information Bottleneck problem was resolved in [1]. We offer insight as to what was aimed to be carried out and emphasize on the deficiencies introduced through such assumptions. We then offer a new model of problem formulation and an accompanying algorithm which could deal with a large amount of previous issues and help reach much more optimal solutions. We offer much mathematical analysis as to why our suggested method is superior to that of [1]'s. In Section IV, we offer some numerical results as another accompanying factor to support our suggested method. We go into details as to what any measure of comparison could entail as a form of our algorithm's superiority. Finally, in Section V, we conclude the paper by reiterating all our suggestions and accomplishments throughout this paper.

II. System Model and Problem Formulation

We assume a Markov chain $U \to X \to Y$ meaning

$$p(Y=y|U=u,X=x)=p(Y=y|X=x) \quad (1)$$

Then, we are seeking a conditional PMF $P_{Y|X}(y|x)$ to attain an optimal tradeoff between quantization rate and information loss about a hidden variable U. We could rewrite this problem in the following manner:

$$\begin{aligned} &min_{P_{Y|X}(y|x)} \quad I(X;Y) \\ &s.t \quad I(U,Y) \geq I_{th} \end{aligned} \quad (2)$$

We will later address how the problem formulated in Eq.(2) represents a non-convex optimization problem and what this classification means for our studies. However, for the time-being, we choose to rewrite Eq.(2) in the format of a Lagrange multiplier problem:

$$min_{P_{Y|X}(y|x)} \quad I(X;Y) - \beta I(U;Y) \quad (3)$$

where $\beta \geq 0$ is assumed to be a Lagrange multiplier connecting the two mutual information functions given to us by the user and we have removed the term $+\beta I_{th}$ from the utility function we aim to minimize as it would be a constant for a fixed value of β.

Furthermore, there are a series of hidden constraints imposed upon the problem which ensure that the final solution $P_{Y|X}(y|x)$ is a probabilistic matrix resulting in:

$$\begin{aligned} &(1) \quad p(y|x) \in [0,1] \forall x \in \mathcal{X}, y \in \mathcal{Y}, \\ &(2) \quad \sum_{y \in Y} p(y|x) = 1, \forall x \in \mathcal{X} \end{aligned} \quad (4)$$

These sets of conditions, while important, could be dealt with in two different manners; one as suggested by [1] and the other as suggested in this paper. As a result, we instead change our focus to the problem formulated in either Eq.(2) or Eq.(3) and only address the conditions in Eq.(4) when needed.

III. Revisiting the Original Information Bottleneck Solution

In this section, we briefly revisit the solution method offered by [1] to exemplify how the problem was tackled previously and so as to offer an in-depth study of the possible limitations of such an algorithm.

To do so, we find it necessary to introduce two types of variables: (1) Free variables which represent the set of variables we are able to choose so as to optimize the overall objective function; in other words, they are the variables achieved by imposing the first order derivative to be equal to 0; (2) Auxiliary variables which help form an iterative relationship with free variables resulting in a more methodical algorithm.

During the remainder of this paper, we will repeatedly refer to these two classes of variables.

A. Problem Formulation

In [1], it was first acknowledged how the problem formulated in Eq.(2) is about minimizing a convex function over a non-convex set. Thus, the overall problem is a non-convex optimization. The proof, while not necessarily difficult, was never fully presented. Here we offer the proof for this observation:

Theorem III.1. *The optimization problem presented in Eq.(2) represents a non-convex optimization.*

The proof for this theorem is presented in Appendix under Theorem III.1. To prove the non-convexity of the optimization, we first show that the functions $I(X;Y)$ and $I(U;Y)$ are both convex functions of $p(y|x)$. Then it could be argued that the set $I(U;Y) \geq I_{th}$ represents a non-convex set while the objective function is still convex in regards to $p(y|x)$ and thus the entire problem in Eq.(2) would be a non-convex optimization. We would advise the reader to study the full proof in the Appendix.

As a direct result of Theorem III.1 , it could be deduced that any solution offered using techniques developed for solving convex optimization problems would be suboptimal. One such suboptimal solution was developed by [1] by reintroducing the original problem in the format of a Lagrange multiplier problem where the goal is to

$$min_{P_{Y|X}(y|x)} \quad I(X;Y) - \beta I(U;Y) - \sum_{x,y} \gamma(x) p(y|x)$$

where once again $\beta > 0$ is assumed to be a Lagrange multiplier connecting the two mutual information functions given to us by the user. Furthermore $\gamma(x)$ represents the constraints imposed by the probabilistic nature of the problem mainly how $\sum_{x,y} p(x,y) = 1$. Moreover, it is assumed that the normalization of the conditional probability matrix is imposed through the solution meaning we will definitely have $\sum_{y_i \in Y} P(y_i|x) = 1, \forall x \in \mathcal{X}$ and we need not worry about such series of conditions for the time being.

Before continuing to the derivation of the problem, we find it necessary to introduce another observation through Eq. (3) which was not previously mentioned in [1].

Theorem III.2. *In the Lagrangian version of the problem as depicted in Eq.(3), we assume $\beta > 1$, otherwise the minimal value of the Lagrangian objective function will be equal to* 0.

The proof to this theorem is gathered in Appendix under Theorem III.2 and turns out to be quite straightforward.

As a direct result of Theorem III.2, from this point on, we will assume that $\beta > 1$ and focus on the answer to this class of problems.

B. An Analysis of the IB Method

In this subsection, we will offer an in-depth study of what [1] suggests in order to solve the problem formulated in Eq.(3). To do so, we will go through the major steps of the analysis in [1], taking into account assumptions made and addressing both pros and cons of iterative algorithms proposed there.

1) Deriving the Optimal $P_{Y|X}(y|x)$: In order to find the optimal $P_{Y|X}(y|x)$ to minimize $\mathcal{L} = I(X;Y) - \beta I(U;Y) - \sum_{x,y} \gamma(x) p(y|x)$, [1] opted to calculate the first order derivative of $\mathcal{L}$ with respect to $\mathbf{P}(\mathbf{y} = \mathbf{y}^*|\mathbf{x} = \mathbf{x}^*)$ for any possible x^* and y^* and have it be equal to 0. In other words, $P_{Y|X}(y|x)$ was chosen as the free variable.

To do so, [1] used the fact that

$$\begin{aligned} p(y) &= \sum_{x \in \mathbf{X}} p(x) p(y|x) \\ &\rightarrow \frac{\partial p(y)}{\partial p(y|x)}|_{x=x^*, y=y^*} = p(x^*) \end{aligned} \tag{5}$$

and the notion that

$$\begin{aligned} p(y|u) &= \sum_{x \in \mathbf{X}} p(x|u) p(y|x) \\ &\rightarrow \frac{\partial p(y|u)}{\partial p(y|x)}|_{x=x^*, y=y^*} = p(x^*|u) \end{aligned} \tag{6}$$

Having made the above assumptions, we would have

$$\begin{aligned} &\frac{\partial \mathcal{L}}{\partial p(y|x)}|_{x=x^*, y=y^*} = 0 \rightarrow \\ &p(x^*)\{\log p(y^*|x^*) + 1 + (\beta - 1) \log p(y^*) + (\beta - 1) \\ &-\beta - \beta \sum_u p(u|x^*) \log \frac{p(y^*) p(u|y^*)}{p(u)}\} - \frac{\gamma(x^*)}{p(x^*)} = 0 \end{aligned} \tag{7}$$

It then follows that

$$\begin{aligned} &\log p(y^*|x^*) = \log p(y^*) + \frac{\gamma(x^*)}{p(x^*)} \\ &+\beta\{D_{KL}\{p(u|x^*)||p(u)\} - D_{KL}\{p(u|x^*)||p(u|y^*)\}\} \\ &\rightarrow p(y^*|x^*) = p(y^*) \\ &\times exp(-\beta\{D_{KL}\{p(u|x^*)||p(u|y^*)\} \\ &-D_{KL}\{p(u|x^*)||p(u)\} - \frac{\gamma(x^*)}{\beta p(x^*)}\}) \\ &\rightarrow p(y^*|x^*) = p(y^*) exp\{-\beta Z(x^*, y^*)\} \end{aligned} \tag{8}$$

where we have used the definition of KL-divergence as $D_{KL}\{\mathbf{P(X)}||\mathbf{Q(X)}\} = \sum_{x \in \mathbf{X}} P(x) \log \frac{P(x)}{Q(x)}$ [8].

Assuming that the original function $\mathcal{L}$ is convex (which we have shown and [1] has mentioned to not be true), then the result as gathered in Eq.(8) will help minimize $\mathcal{L}$.

However, there is still another issue which needs to be addressed. $p(y^*|x^*)$s as gathered in Eq.(8) do not necessarily indicate a set of probability variables; while they are always greater than 0, there is no guarantee that they will remain below 1 or even that the sum of all of its possible values follow the marginal property (both defined in Eq.(4)).

[1] suggested that in order to deal with this issue, we would redefine the optimal $p(y^*|x^*)$ as

$$p(y^*|x^*) = \frac{p(y^*) exp\{-\beta Z(x^*, y^*)\}}{\sum_{y^{**}} p(y^{**}) exp\{-\beta Z(x^*, y^{**})\}} \tag{9}$$

By doing so, [1] has guaranteed that $\sum_{y \in Y} p(y|x) = 1$ and that $p(y|x) \leq 1$. However, normalizing $p(y^*|x^*)$ as seen in Eq.(9), takes away from the concept of an optimal $p(y^*|x^*)$ in the first place. Thus, even if convexity of $\mathcal{L}$ were guaranteed, by the above imposition, the calculated $p(y^*|x^*)$ is no longer the optimal minimizer we were looking for. Later on, we will offer further details as to why such an argument is important and quite problematic.

We could thus conclude that in the calculation of the optimal $p(y^*|x^*)$ as described in [1] much has been left to desire.

2) Derivating the optimal $P_Y(y)$: In [1], once the optimal free variable $p(y^*|x^*)$ was calculated, it was assumed that we would treat $P_Y(y)$ as auxiliary variables and fix

$$p(y^*) = \sum_{x \in \mathbf{X}} p(x)p(y^*|x), \forall y^* \in \mathcal{Y} \tag{10}$$

Such an assumption would guarantee another series of conditions that all marginal probability distributions should satisfy. Finally, it was suggested that by repeatedly fixing $p(y^*)$ and then finding the optimal $p(y^*|x^*)$ and vice versa, we could reach a point of convergence where the overall function $\mathcal{L}$ is minimized.

We have already shown that for a fixed $p(y^*)$, $p(y^*|x^*)$s as formulated in Eq.(9) are not necessarily optimal (they are not moving in the minimization direction). In this section, we show that for a fixed set of $p(y^*|x^*)$s, $p(y^*)$s as formulated in Eq.(10) do not move in minimization direction either.

To show this, we assume two different functions $\mathcal{F} = E_{x,y}[\log \frac{P_{Y|X}(y|x)}{Q_Y(y)}]$ and $\mathcal{G} = E_{x,y}[\log \frac{P_{Y|X}(y|x)}{\sum_{x \in \mathbf{X}} P_X(x)P_{Y|X}(y|x)}]$.

Note that in the function of $\mathcal{F}$, $Q()$ is a legitimate probability measure on $\mathcal{Y}$, but not necessary satisfying Bayes rule as in Eq.(10). This is because we are attempting to show some relationship between $\mathcal{F}$ and $\mathcal{G}$ for a more general setting. It follows that

$$\begin{aligned}\mathcal{F} - \mathcal{G} &= \sum_{x,y} P_X(x)P_{Y|X}(y|x) \log \frac{\sum_x P_X(x)P_{Y|X}(y|x)}{Q_Y(y)} \\ &\geq \sum_{x,y} P_X(x)P_{Y|X}(y|x)\{1 - \frac{Q_Y(y)}{\sum_x P_X(x)P_{Y|X}}\} \\ &= 1 - 1 = 0\end{aligned} \tag{11}$$

where in the penultimate line we have used the log-inequality $\log x \geq 1 - \frac{1}{x}$ with $=$ only appearing when $x = 1$. (which in this scenario is equivalent to the condition that $Q_Y(y) = \sum_x P_X(x)P_{Y|X}, \forall y \in \mathcal{Y}$)

It follows that if we define

$$\mathcal{K} = E_{x,y}[\log \frac{P_{Y|X}(y|x)}{Q_Y(y)}] - \beta E_{u,y}[\log \frac{P_{Y|U}(y|u)}{Q_Y(y)}] \tag{12}$$

and

$$\begin{aligned}\mathcal{M} = E_{x,y}[\log \frac{P_{Y|X}(y|x)}{\sum_{x \in \mathbf{X}} P_X(x)P_{Y|X}(y|x)}] \\ -\beta E_{u,y}[\log \frac{P_{Y|U}(y|u)}{\sum_{x \in \mathbf{X}} P_X(x)P_{Y|X}(y|x)}]\end{aligned} \tag{13}$$

Then, we will not be able to derive much about the difference between $\mathcal{K}$ and $\mathcal{M}$ as

$$\begin{aligned}\mathcal{K} - \mathcal{L} = \sum_{x,y} P_X(x)P_{Y|X}(y|x) \log \frac{\sum_x P_X(x)P_{Y|X}(y|x)}{Q_Y(y)} \\ -\beta \sum_{u,x,y} P_X(x)P_{U|X}(u|x)P_{Y|X}(y|x) \\ \log \frac{\sum_x P_X(x)P_{Y|X}(y|x)}{Q_Y(y)}\end{aligned} \tag{14}$$

where the first sum is always positive and the second always negative. As a result, based on the value of β the overall difference may be either positive or negative. If we assume $\beta << 1$, then the overall difference will be positive and as a result fixing $Q_Y(y) = \sum_x P_X(x)P_{Y|X}(y|x)$ will minimize the overall difference. On the other hand, for $\beta >> 1$, the overall difference will be negative and as a result fixing $Q_Y(y) = \sum_x P_X(x)P_{Y|X}(y|x)$ will maximize the overall difference.

Overall, it could be deduced that fixing $Q_Y(y) = \sum_x P_X(x)P_{Y|X}(y|x)$ will not necessarily result in minimizing $\mathcal{L}$ for a fixed $P_{Y|X}(y|x)$ and is thus not optimal.

3) Overall Conclusion of the Original IB Method: Overall, one of the primary contributions of [1] was to offer two closed-form formulas by whose iterations and many choices of starting variables, it could be shown that a final value of $\mathcal{L}$ could be reached. However, as was mentioned in [1], this overall value is not necessarily the minimum we were looking for. In this section, we went one step further and demonstrated how this non-optimality is not only due to the non-convexity of the original problem (as [1] had pointed out), but also due to the method [1] has offered.

To put it simply, neither of the two Equations (9) and (10) as derived by [1] necessarily move the overall objective function $\mathcal{L}$ in a decreasing fashion.

4) An Idea to Implement: By revisiting the original formulation of $\mathcal{L}$ it could be witnessed that for a fixed $I(X;Y)$, further increasing $I(U;Y)$ would result in the minimization of the overall $\mathcal{L}$.

As a result, we suggest treating $P_{Y|U}(y|u)$ as a completely independent set of variables. We will simply introduce a series of conditions required to be satisfied to guarantee the linear relationship between $P_Y(y)$, $P_{Y|X}(y|x)$ and $P_{Y|U}(y|u)$. Consequently, we propose to turn the objective function $\mathcal{L}$ as following:

$$\mathcal{L} = E_{x,y}[\log \frac{P_{Y|X}(y|x)}{Q_Y(y)}] - \beta E_{u,y}[\log \frac{P_{Y|U}(y|u)}{Q_Y(y)}] \tag{15}$$

Such an idea has 2 very important advantages in our optimization problem:

1: In Eq.(15), it could be seen that the overall $\mathcal{L}$ is a convex function of $P_{Y|X}(y|x)$ (check the proof for Theorem III.1) and thus minimization overall $P_{Y|X}(y|x)$ alone is calculable.

2: Assuming a fixed $P_{Y|X}(y|x)$ and $P_Y(y)$, we need to show that there exists a class of variables $P_{Y|U}(y|u)$ which will help minimize the overall $\mathcal{L}$. By utilizing the same idea introduced in [3] and assuming that $P_Y(y)$ and $P_{Y|X}(y|x)$ are fixed and introducing $\mathcal{F}' = E_{u,y}[\log \frac{Q_{Y|U}(y|u)}{P_Y(y)}]$ and $\mathcal{G}' = E_{u,y}[\log \frac{W_{Y|U}(y|u)}{P_Y(y)}]$ where $W_{Y|U}(y|u) = \frac{P_Y(y)P_{U|Y}(u|y)}{\sum_{y \in \mathbf{Y}} P_Y(y)P_{U|Y}(u|y)}$, we can write

$$\begin{aligned}\mathcal{G}' - \mathcal{F}' = \sum_{u,y} P_Y(y)P_{U|Y}(u|y) \\ \log \frac{P_Y(y)P_{U|Y}(u|y)}{Q_{Y|U}(y|u)\sum_{y^*} P_Y(y^*)P_{U|Y}(u|y^*)} \\ \geq \sum_{u,y} P_{U|Y}(u|y)P_Y(y) - \sum_{u,y} Q_{Y|U}(y|u)P_U(u) \\ = 1 - 1 = 0\end{aligned} \tag{16}$$

where we have once again utilized the log inequality and $=$ only holds true when

$$Q_{Y|U}(y|u) = \frac{P_Y(y)P_{U|Y}(u|y)}{\sum_{y\in\mathbf{Y}} P_Y(y)P_{U|Y}(u|y)} = \frac{\sum_x P_{X|U}(x|u)P_{Y|X}(y|x)}{\sum_{x,y} P_{X|U}(x|u)P_{Y|X}(y|x)} \quad (17)$$

where we have simplified the first line using some simple mathematical calculations. It follows that satisfying Eq.(17) results in maximizing $\mathcal{F}'$ and thus minimizing $\mathcal{L}$ for a fixed $P_{Y|X}(y|x)$ and $P_Y(y)$. Interestingly, this condition is a direct result of imposing an extension of the Bayes rules on variables $P_{Y|U}(y|u)$ and $P_{Y|X}(y|x)$. As a result, we choose to select a third class of variables namely $P_{Y|U}(y|u)$ as another set of auxiliary variables to be updated alternatively together with $P_Y(y)$, and $P_{Y|X}(y|x)$.

As was witnessed, the biggest fall-back of [1]'s suggestions appeared when an attempt was made to ensure

that the Bayes conditions between $P(Y)$ and $P(Y|X)$ and that between $P(Y|U)$ and $P(Y|X)$ underlying the Markov Chain $U \to X \to Y$ are reinforced with respect to only one free set of variables $P(Y|X)$. In order to deal with these problems, we suggest the use of Augmented Lagrange Multiplier (ALM) Method and more specifically the Alternating Direction Method of Multipliers (ADMM) which we will discuss in the follwing section.

C. Introducing Penalty (cost) Functions

After careful observation of the issues introduced in the previous subsection, we could deduce that most of the deficiencies of [1] come from a restrictive manner of dealing with constrained non-convex optimization problems. In both calculations, we have sacrificed the optimal direction of reaching a solution for making certain a constraint is met at every step of the iteration. It would then make sense to try and find a solution which can satisfy an acceptable threshold of constraints at every step of the iteration while not diverging from the optimal path of reaching a solution. It turns out such a solution could be developed by introducing a series of cost functions.

D. Constrained Optimization using Augmented Lagrange Multipliers

As mentioned previously, the biggest issue with the IB solution lays in the strict nature of a constraint resulting in many inconsistencies while developing an iterative algorithm. More specifically, we either allow the constraints to be fully broken so as to develop a straightforward solution (the ideal $p(y^*|x^*)$ calculated in Eq.(8)) or we impose the constraints so rigidly that they result in the a misdirection in the solution as was the case for imposing the marginal property in Eq.(10).

A method of dealing with such scenarios was addressed by the Augmented Lagrange Multiplier method [2] which allows a level of flexibility from such constraints but also introduces a penalty function for such diversions. As a result, by controlling the value of the penalty coefficient, we will be able to both account for possible diversions and also keep them under a specific threshold. Furthermore, by imposing a large penalty function, we will be able to virtually find answers when the constraints are not broken at all (We will discuss this property in full details while observing the numerical results).

A secondary source of problems in developing [1]'s method could be traced back to forming an iterative relationship between only 2 classes of variables (namely $P_{Y|X}(y|x)$ and $P_Y(y)$s. In [1], by simply fixing one of the classes, we would try and find the other optimal class of variables. Such a solution would impose further weight and constraints over one of the variable classes. For example in the problem at hand as exemplified in Eq.(3), if we fix $P_Y(y)$, the optimal value of $P_{Y|X}(y|x)$ needs to account for traces of a secondary class of variables tentatively known as $P_{Y|U}(y|u)$'s as demonstrated through Eq.(6). We could argue that by further introducing new classes of variables, we could cut then on such variable interconnections thus dealing with simpler problems. As a result, from this point on, we assume we would like to optimize $\mathcal{L}$ over 3 sets of variable classes $P_Y(y)$, $P_{Y|X}(y|x)$ and $P_{Y|U}(y|u)$.

Next, we will go into details as to how such penalty coefficients could be accounted for and applied to our problem at hand.

1) Introducing Penalties: To introduce the penalty coefficients in our constrained optimization problem, we first list all possible constraints in detail assuming we are searching over all sets of variables $P_Y(y), P(Y|X)$ and $P_{Y|U}(y|u)$s. We break these constraints into those of an inequality nature:

$$\begin{aligned} 0 \le p(y|x) \le 1, \forall x \in \mathcal{X}, \forall y \in \mathcal{Y} \\ 0 \le p(y|u) \le 1, \forall u \in \mathcal{U}, \forall y \in \mathcal{Y} \\ 0 \le p(y) \le 1, \forall y \in \mathcal{Y} \end{aligned} \quad (18)$$

and those of an equality nature such as:

$$\begin{aligned} \sum_x p(x)p(y|x) = p(y), \quad \forall y \in \mathcal{Y} \\ p(y|u) = \frac{\sum_x p(x|u)p(y|x)}{\sum_{x,y} p(x|u)p(y|x)}, \quad \forall y \in \mathcal{Y}, \forall u \in \mathcal{U} \\ \sum_y p(y|x) = 1, \quad \forall x \in \mathcal{X} \\ \sum_y p(y|u) = 1, \quad \forall u \in \mathcal{U}, \quad \sum_y p(y) = 1 \end{aligned} \quad (19)$$

Note: It should be noted that all the variables $p(y|u)$ and $p(y)$ are linear functions of the original variables $p(y|x)$ and thus, the problem is still that of an optimization over the matrix $P_{Y|X}(y|x)$ with all the constraints over the same matrix. However, in order to further simplify the calculations and use the results from other works specifically ADMM [9], we treat $p(y|u)$ and $p(y)$ as separate variables and then use the set of constraints gathered in Eq.(18) and Eq.(19) as new constraints within theses variables.

E. New Problem Formulation using Penalties

To offer the new formulation, we first need to introduce some definitions:

1. Any $\lambda, \lambda', \lambda'', \lambda''', \lambda''''$ represents the Lagrange multipliers defined over their corresponding set of equality constraints and their indices simply reveal over which variables they are defined.
2. Any μ, μ' represents the Lagrange multipliers defined over their corresponding set of inequality constraints and their indices simply reveal over which variables they are defined.
3. A penalty scalar c is defined.

Using the concept of ALM over constrained sets, we could write the overall new utility function in the format of:

$$\mathcal{L}_{\mathbf{c}} = \sum_{x,y} p(y|x) \log p(y|x) - \sum_{y} p(y) \log p(y)$$
$$+ \sum_{y} \lambda_y \quad \{p(y) - \sum_{x} p(x)p(y|x)\}$$
$$+ \sum_{u,y} \lambda'_{u,y} \quad \{p(y|u) - \frac{\sum_x p(x|u)p(y|x)}{\sum_{x,y} p(x|u)p(y|x)}\}$$

$$+\lambda'' \quad \{\sum_{y} p(y) - 1\} + \sum_{u} \lambda'''_u \quad \{\sum_{y} p(y|u) - 1\}$$
$$+ \sum_{x} \lambda''''_x \quad \{\sum_{y} p(y|x) - 1\}$$
$$+ \frac{c}{2} \sum_{y} \{p(y) - \sum_{x} p(x)p(y|x)\}^2$$
$$+ \frac{c}{2} \sum_{u,y} \{p(y|u) - \frac{\sum_x p(x|u)p(y|x)}{\sum_{x,y} p(x|u)p(y|x)}\}^2$$
$$+ \frac{c}{2} \sum_{y} \{p(y) - 1\}^2$$
$$+ \frac{c}{2} \sum_{u} \{\sum_{y} p(y|u) - 1\}^2 + \frac{c}{2} \sum_{x} \{\sum_{y} p(y|x) - 1\}^2$$
$$+ \frac{1}{2c} \sum_{x,y} \{max(0, \mu_{xy} - cp(y|x))^2 - \mu_{xy}^2\}$$
$$+ \frac{1}{2c} \sum_{u,y} \{max(0, \mu_{uy} - cp(y|u))^2 - \mu_{uy}^2\}$$
$$+ \frac{1}{2c} \sum_{y} \{max(0, \mu_{y} - cp(y))^2 - \mu_{y}^2\}$$
$$+ \frac{1}{2c} \sum_{x,y} \{max(0, \mu'_{xy} - c + cp(y|x))^2 - {\mu'_{xy}}^2\}$$

$$+ \frac{1}{2c} \sum_{u,y} \{max(0, \mu'_{uy} - c + cp(y|u))^2 - {\mu'_{uy}}^2\}$$
$$+ \frac{1}{2c} \sum_{y} \{max(0, \mu'_{y} - c + cp(y))^2 - {\mu'_{y}}^2\} \tag{20}$$

We need to address two new additions in the previous steps:

(1) There is a use of max function in the formulation of Eq.(20) which is due to the fact that we have chosen to rewrite the inequality constraints in the format of equality constraints using a positive variable. We then find the optimal value for such positive variables to minimize the overall utility function. This then in turn results in choosing a value between the optimal and 0, then if the optimal is positive, the optimal is chosen and otherwise, 0 is chosen which is the optimal allowed value [2].

(2) We have chosen a cost scalar c which controls the level of allowed invasion of constraints over the course of optimization [2].

F. Solution Derivation

We find the optimal set of answers $P_{Y|X}(y|x)$, $P_Y(y)$ and $P_{Y|U}(y|u)$ over which the function $\mathcal{L}_{\mathbf{c}}$ is minimized. We could carry this out in 3 different ways:

(1) We could solve the problem using the original Augmented Lagrange Multiplier Method as described in [2]. To do so, we will need to carry out the gradient descent method over a total of $|Y|$ variables (to account for all $p(y)$s) plus $|X| \times |Y|$ variables (to account for all $p(y|x)$s) plus $|U| \times |Y|$ variables (to account for all $p(y|u)$s) at the same time. Such work, while computationally doable, is (1) very time-consuming and (2) does not offer a high convergence rate. This limit in convergence is due to the fact that we are trying to minimize an objective function $\mathcal{L}$ over a large number of variables each of whom impose their own set of constraints (λ 's and μs) and all of which need to be at least partially satisfied.

(2) We could solve this problem by optimizing $\mathcal{L}$ over different variable classes separately. To do so, we assume all $P_{Y|X}(y|x)$ and $P_{Y|U}(y|u)$s are fixed and then try to minimize $\mathcal{L}$ over all possible $P_Y(y)$s. We then fix $P_Y(y)$ and $P_{Y|X}(y|x)$ and try to optimize $\mathcal{L}$ over all $P_{Y|U}(y|u)$s and so forth. If we continue to iteratively optimize $\mathcal{L}$ over such sets, we could then hope that we will end up at a desirable minimum value for $\mathcal{L}$. By doing so, we will be able to cut down on the level of complications arising from optimization over a large number of variables as was the case in the original Augmented Lagrange Multiplier Method (ALM) and thus multiply the rate of convergence. This method is widely referred to as Alternating Direction Method of Multipliers (ADMM) and has been used in many recent studies to discuss non-convex optimization problems (as is the case in our problem).

(3) We could opt to use randomized ADMM [10] which is simply an extension of the original ADMM method as described in (2) where at every step of optimization, the order of optimization over different sets of varibale classes $P_{Y|X}(y|x)$, $P_{Y|U}(y|u)$ and $P_Y(y)$ is randomly selected. In other words, at every step, 1 of the 6 possible permutations of these variables is randomly chosen and then ADMM is carried out. While never mathematically proven, this method has been shown to offer a better optimization result through many practical implementations. [10], [11]

Through the remainder of this paper, we choose to further focus on ADMM and will suffice to present results of solving the same problem using ALM as a means of comparison between the two methods.

Note: It is important to note that in both ALM and ADMM, optimization is carried out through gradient descent. The only difference is that by using ADMM, we are breaking down a large scale ALM problem to a a number of smaller scale ALM problems thus raising our chances of a converging solution.

G. An In-depth Study of the Solution

In this section, we offer the reader further insight into how a method like ADMM could help us sense the subtle deficiencies of the method presented in [1]. As was the case in [1], we assume that the function formulated in Eq.(20) is convex with respect to each of the 3 possible variable classes. We then try to calculate the first order derivative of new $\mathcal{L}$ as defined in Eq.(20) with respect to each variable class and have it be equal to 0 to see how each of these 3 sets of equations hold their own against what was happening in [1]'s solution.

Note: At this part of the paper, we find it necessary to discuss the concept of derivation of a function W defined as $W = \max(a(\omega), b(\omega))$; we assume that we have access to the latest value of function W. If for that value, a is the maximal, we calculate $\frac{\partial a}{\partial \omega}$ and otherwise, we calculate $\frac{\partial b}{\partial \omega}$. In better words, we choose to calculate one sided first order derivations over the W function.

Using the definition above, we begin calculating the first order derivatives of $\mathcal{L}$ with respect to all possible variables:

$$\frac{\partial \mathcal{L}_{\mathbf{c}}}{\partial p(y|x)}|_{x=x^*,y=y^*} = 0 \rightarrow$$

$$\frac{p(x^*)}{\log 2}\{\log p(y^*|x^*) + 1\} + \lambda_{y^*}\{-p(x^*)\} + \lambda''''_{x^*}$$

$$\sum_u \lambda'_{u,y^*} p(x^*|u)\{-1 + \frac{\sum_x p(x|u)p(y^*|x)}{\sum_{x,y} p(x|u)p(y|x)}\}$$

$$+c\sum_{u,y} p(x^*|u)\{-1 + \frac{\sum_x p(x|u)p(y^*|x)}{\sum_{x,y} p(x|u)p(y|x)}\}\times$$

$$\{p(y|u) - \frac{\sum_x p(x|u)p(y|x)}{\sum_{x,y} p(x|u)p(y|x)}\}$$

$$+cp(x^*)\{\sum_x p(x)p(y^*|x) - p(y^*)\} + c\{\sum_y p(y|x^*) - 1\}$$

$$+\begin{cases} cp(y^*|x^*) - \mu_{x^*y^*} & \mu_{x^*y^*} \geq cp(y^*|x^*) \\ 0 & \mu_{x^*y^*} < cp(y^*|x^*) \end{cases}$$

$$+\begin{cases} cp(y^*|x^*) + \mu'_{x^*y^*} - c & \mu'_{x^*y^*} \geq c - cp(y^*|x^*) \\ 0 & \mu'_{x^*y^*} < c - cp(y^*|x^*) \end{cases} = 0 \tag{21}$$

Eq.(21) represents a relationship between $p(y^*)$ and $p(y^*|x^*)$ in the same manner that Eq.(6) did for the original information bottleneck method [1]. However, here Eq.(21) is just one of three sets of equalities that need to hold true.

We now follow the same manner for the other two variables:

$$\frac{\partial \mathcal{L}_{\mathbf{c}}}{\partial p(y|u)}|u = u^*, y = y^* = 0 \rightarrow$$

$$\lambda'_{y^*}\{-p(u^*)\} + \lambda'''_{u^*} + \lambda'_{u^*,y^*}$$

$$+c\sum_{u,y} p(y|u) - \frac{\sum_x p(x|u)p(y|x)}{\sum_{x,y} p(x|u)p(y|x)}$$

$$+cp(u^*)\{\sum_u p(u)p(y^*|u) - p(y^*)\} + c\{\sum_y p(y|u^*) - 1\}$$

$$+\begin{cases} cp(y^*|u^*) - \mu_{u^*y^*} & \mu_{u^*y^*} \geq cp(y^*|u^*) \\ 0 & \mu_{u^*y^*} < cp(y^*|u^*) \end{cases}$$

$$+\begin{cases} cp(y^*|u^*) + \mu'_{u^*y^*} - c & \mu'_{u^*y^*} \geq c - cp(y^*|u^*) \\ 0 & \mu'_{u^*y^*} < c - cp(y^*|u^*) \end{cases}$$

$$= 0 \tag{22}$$

Eq.(22) represents a relationship between $p(y^*)$ and $p(y^*|u^*)$ in the same manner as Eq.(21). However, we can witness that this relationship is quite diverse in nature where for different values of u^*, y^*, the relationship between variables (while still always linear) is prone to change.

Finally, we have:

$$\frac{\partial \mathcal{L}_{\mathbf{c}}}{\partial p(y)}|_{y=y^*} = 0 \rightarrow$$

$$-\{1 + \log p(y^*)\} + \lambda_{y^*} + \lambda'_{y^*} + \lambda''$$

$$+c\{p(y^*) - \sum_x p(x)p(y^*|x)\} + c\{p(y^*) - \sum_u p(u)p(y^*|u)\}$$

$$+c\{p(y^*) - 1\} + \begin{cases} cp(y^*) - \mu_{y^*} & \mu_{y^*} \geq cp(y^*) \\ 0 & \mu_{y^*} < cp(y^*) \end{cases}$$

$$+\begin{cases} cp(y^*) + \mu'_{y^*} - c & \mu'_{y^*} \geq c - cp(y^*) \\ 0 & \mu'_{y^*} < c - cp(y^*) \end{cases}$$

$$= 0 \tag{23}$$

Once again, we could see that the relationship between $p(y^*), p(y^*|x^*), p(y^*|u^*)$ could drastically change seeing as how based upon the margins, the sign of $\log p(y^*)$ might change; thus changing the entire format of Eq.(23).

We then follow the same intuition as that of the Information Bottleneck (IB) method by using each of Eq.(19,22) and Eq.(23) to update the values of $p(y^*|x^*)$, $p(y^*|u^*)$ and $p(y^*)$ respectively. Furthermore, we continuously update the Lagrange multipliers (based upon whether they represent an equality or an inequality constraint) as follows:

$$\lambda_{y^*}{}^{(t+1)} = \lambda_{y^*}^{(t)} + c\{p^{(t)}(y^*) - \sum_x p(x)p^{(t)}(y^*|x)\}$$

$$\lambda'_{u^*,y^*}{}^{(t+1)} = \lambda'_{u^*,y^*}{}^{(t)} +$$

$$c\{p^{(t)}(y^*|u^*) - \frac{\sum_x p(x|u)p(y|x)}{\sum_{x,y} p(x|u)p(y|x)}\}$$

$$\lambda''^{(t+1)} = \lambda''^{(t)} + c\{-1 + \sum_y p^{(t)}(y)\}$$
$$\lambda'''^{(t+1)}_{u^*} = \lambda'''^{(t)}_{u^*} + c\{-1 + \sum_y p^{(t)}(y|u)\}$$
$$\lambda''''^{(t+1)}_{x^*} = \lambda''''^{(t)}_{x^*} + c\{-1 + \sum_y p^{(t)}(y|x)\}$$
$$\mu^{(t+1)}_{x^*y^*} = \frac{\mu^{(t)}_{x^*y^*}}{2}, {\mu'_{x^*y^*}}^{(t+1)} = \frac{{\mu'_{x^*y^*}}^{(t)}}{2}$$
$$\mu^{(t+1)}_{u^*y^*} = \frac{\mu^{(t)}_{u^*y^*}}{2}, {\mu'_{u^*y^*}}^{(t+1)} = \frac{{\mu'_{u^*y^*}}^{(t)}}{2}$$
$$\mu^{(t+1)}_{y^*} = \frac{\mu^{(t)}_{y^*}}{2}, {\mu'_{y^*}}^{(t+1)} = \frac{\mu^{(t)}_{y^*}}{2} \quad (24)$$

Finally, we follow these steps as an algorithm of solving the Information Bottleneck problem:

Step 0: Choose random initial values for $p^{(0)}(y^*|x^*)$,$p^{(0)}(y^*|u^*)$ and $p^{(0)}(y^*)$ and all $\lambda^{(0)}$s and $\mu^{(0)}$s.
Step 1: Run Gradient Descent Algorithm to find $p^{(t+1)}(y^*|x^*)$ which minimizes $\mathcal{L}_\rfloor$ for fixed values of $p^{(t)}(y^*|u^*)$ and $p^{(t)}(y^*)$ and all $\lambda^{(t)}$s and $\mu^{(t)}$s.
Step 2: Run Gradient Descent Algorithm to find $p^{(t+1)}(y^*|u^*)$ which minimizes $\mathcal{L}_\rfloor$ for fixed values of $p^{(t)}(y^*|x^*)$ and $p^{(t)}(y^*)$ and all $\lambda^{(t)}$s and $\mu^{(t)}$s.
Step 3: Run Gradient Descent Algorithm to find $p^{(t+1)}(y^*)$ which minimizes $\mathcal{L}_\rfloor$ for fixed values of $p^{(t)}(y^*|x^*)$,$p^{(t)}(y^*|u^*)$ and all $\lambda^{(t)}$s and $\mu^{(t)}$s.
Step 4: Use Eq.(24) to update $\lambda^{(t+1)}$s and $\mu^{(t+1)}$s.
Step 5: Go to step 2 unless all variables from steps t and $t+1$ are equal.

IV. Numerical Results and Comparison

In this section, we run 3 different algorithms upon 3 different settings ($U \to X$ probability distributions channels) assuming that in all cases $|X| = |U| = 3, |Y| = 2$. These channels can be found in Table I. For all cases, we have assumed that

$$\mathbf{P}(\mathbf{X}|\mathbf{U}) = \begin{bmatrix} 0.3 & 0.1 & 0.6 \\ 0.8 & 0.1 & 0.1 \\ 0.5 & 0.4 & 0.1 \end{bmatrix}$$

Table I
$\mathbf{P}(\mathbf{U})$ settings in each example

1	2
$\begin{bmatrix} 0.3 & 0.4 & 0.3 \end{bmatrix}$	$\begin{bmatrix} 0.2 & 0.4 & 0.4 \end{bmatrix}$

The algorithms we opt to use are the original IB method [1], our proposed ADMM based algorithm and a randomized ADMM method.

A. Grounds for Comparison

In order to fairly compare the algorithms, we take note of how the IB method requires a given constant $\beta > 1$.

As a result, our first measure of comparison is to plot the final calculated value of the objective function $\mathcal{L}$ versus

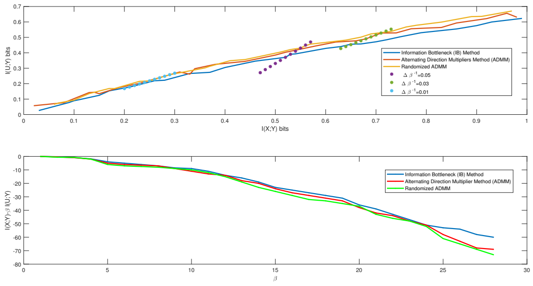


Figure 1. Comparison of (1) $I(U;Y)$ for a fixed $I(X;Y)$ and (2) $\mathcal{L}$ for Example 1

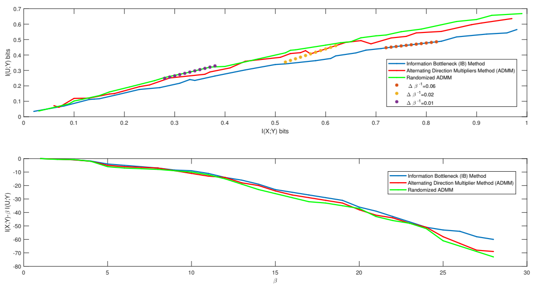


Figure 2. Comparison of (1) $I(U;Y)$ for a fixed $I(X;Y)$ and (2) $\mathcal{L}$ for Example 2

differing values of $\beta > 1$. Then, one method of comparing the efficiency of each method is to see which one of them offers a minimal $\mathcal{L}$ and is thus the optimal in our scenario. By doing so, we are inadvertently choosing the probability distribution matrix which helps minimize the function gathered in Eq.(3). These results are gathered in part (2) of Figures 1 and 2 respectively.

Furthermore, we opt to take note of another measure of comparison as well; we compare different values of $I(U;Y)$ for a fixed value of $I(X;Y)$. By doing so, we are somehow moving away from the Lagrange multiplier nature of the problem and instead focusing on the more constrained optimization nature of it instead (as better formulated in Eq.(2). Then, one method of comparing the efficiency of each method is to see which one of them offers the maximal $\mathcal{L}$ and is thus the optimal choice in our scenario. The results are gathered in part (1) of Figures 1 and 2 respectively.

B. Final Results

As can be witnessed, in both series of results, the optimal value is achieved through the use of randomized ADMM, followed by normal ADMM followed by the original IB method. As expected, there are times when two or even all 3 of the algorithms result in the same overall objective function however there are many times when one overpowers the other 2.

Finally, we would like to offer further insight as to why we have chosen to plot graphs representing $I(U;Y)$ vs $I(X;Y)$ at all. It might seems like just by plotting $\mathcal{L}$ for different values of $\beta > 1$, one might be able to judge the efficiency of randomized ADMM over all other methods.

While that may be true, we are more interested in the relationship between these two variables.

[1] previously discussed how in an optimal setting (if the problem were convex), the optimal solution of the problem formulated in Eq.(3) would satisfy:

$$\frac{\partial I(U;Y)}{\partial I(X;Y)} = \frac{1}{\beta} \tag{25}$$

This is a direct conclusion from optimizing $\mathcal{L}$ by calculating its first derivative and putting it equal to 0. Using this intuition, we can now compare our solution with the optimal we are hoping to achieve for every possible solution.

We refer the reader to part (1) of Figures 1 and 2. Optimally speaking, we hope to find the part of the figure whose slope is the closest to $\frac{1}{\beta}$ for a given β. The slopes indicated on the figure represent the closest slopes we could get to $\frac{1}{\beta} = \frac{1}{2} = 0.5$. As can be seen, the least amount of difference between the two slopes $\Delta\frac{1}{\beta}$ belongs to randomized ADMM followed by normal ADMM followed by IB. This offers us another insight as to why randomized ADMM is the optimal solution to our problem at hand seeing as how it trails the optimal setting the closest.

C. Convergence versus Accuracy

As a means of better understanding how Augmented Lagrange Multiplier whether it be ALM or ADMM improves our chances of a solution, we decided to plot a series of figures indicating how different quality factors of the problem change based upon the value of the penalty coefficient (indicated as c during this paper). The final results turned out to be quite interesting and well worth a deeper study.

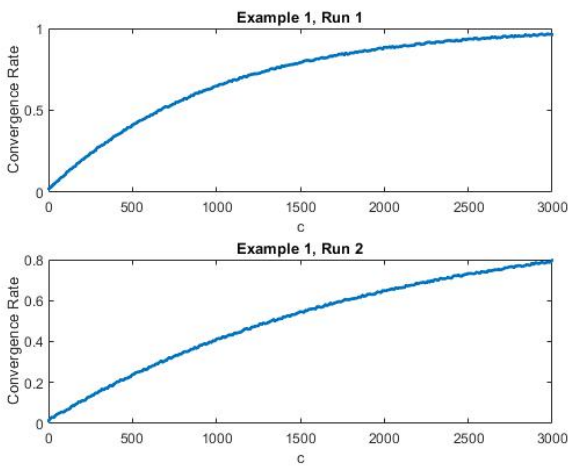


Figure 3. Convergence Rate vs c for Example 1 in (1) Run 1 and (2) Run 2

Parts (1) and (2) of Figure 3 demonstrate the convergence rate of the ADMM solution versus the value of c for Example 1 in 2 runs of $10,000$ repetitions. As can be seen, as a general rule, as c grows larger, the convergence rate grows higher and higher until for some large enough value of c, the method is almost always converging. Unfortunately, this large value of c is very case-dependent, as under the same circumstances and with simply different starting points (Step 0 of the algorithm) it could drastically change.

On the other hand, we have Figure 4 which demonstrates the accuracy of the solution we end up with at the end of running our algorithm. Such accuracy is calculated through the measure $\Delta\frac{1}{\beta}$ which was fully described previously. An accuracy of 1 means $\Delta\frac{1}{\beta} \to 0$ and an accuracy of almost zero refers to when $\Delta\frac{1}{\beta} \to max(\frac{1}{\beta}, 1-\frac{1}{\beta})$ where the 2 extremes of the max refer to when the closest slope to the desired β is infinity and 1 respectively. The resulting graph will be in the following format.

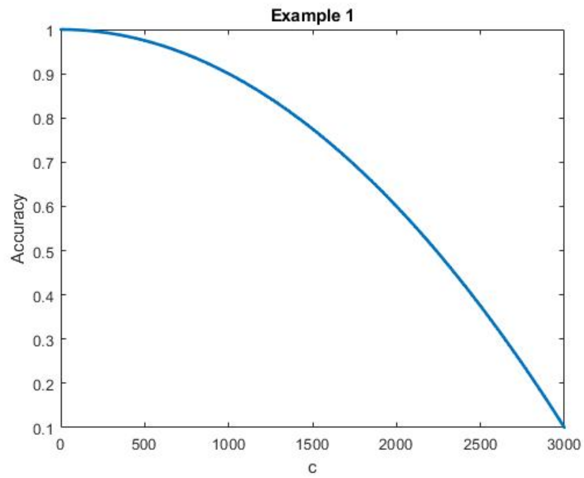


Figure 4. Accuracy Rate vs c for Example 1

It's quite interesting to see that the 2 measures of quality act completely opposite one another as c increases. In order to explain this phenomenon, we need to remind the reader of the meaning behind Eq.(22). By allowing penalty coefficients in the form of quadratic functions, we are simplifying some of the more non-convex elements of the original objective function. It then follows that if c grows larger, the quadratic elements of the new objective function will become more prominent and could direct the entire objective function into a completely convex nature thus resulting in a very high convergence rate. However, if we keep increasing c with no regards to the original objective function, we will no longer be solving an optimization over $\mathcal{L}$ but rather a new objective mainly defined over c and thus while an answer will almost certainly exist (if c is large enough), it will start to diverge drastically from the desired optimization problem (a very low accuracy). [6]

Finally, we would like to investigate if an actual lower bound for c could be found to ensure that the overall objective function will converge. So far, in the 2 examples provided above, it could be witnessed that such a lower bound might not be truly calculable (as the threshold changes for the same problem settings with only different starting points).

[7] discusses a series of sufficient conditions whose satisfaction would result in the absolute convergence of a non-convex optimization problem. Namely, [7] introduces the concept of Lipschitz continuity over each select set of variable class and shows how if such a condition is granted,

convergence is guaranteed. To do so, [7] introduces the concept of Lipschitz Differentiability which needs to hold true over functions of each separate variable set. Next, we will show that unfortunately such a series of conditions do not hold true in our scenario.

Unfortunately, it could be shown that such sufficient conditions do not hold true over our class of optimization problem and thus we could not use the results from [7] to make any further conclusions.

V. Conclusions

In this paper, we revisited the Information Bottleneck problem and the widely used method of solving it. We noted the simplicity of this method but also demonstrated its weaknesses. We then offered a new method of solving the same problem which takes care of such weaknesses while sacrificing the simplicity of the original in exchange for making certain the constraint are all held and the final solutions are superior to those of the original method.

References

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 1999, pp. 368–377.

[2] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[3] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, July 1972.

[4] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, January 1972.

[5] A. Ghassami, S. Khodadadian, and N. Kiyavash, "Fairness in supervised learning: An information theoretic approach," *CoRR*, vol. abs/1801.04378, 2018. [Online]. Available: http://arxiv.org/abs/1801.04378

[6] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," in *2015 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015 - Proceedings*, ser. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. Institute of Electrical and Electronics Engineers Inc., 8 2015, pp. 3836–3840.

[7] Y. Wang, W. Yin, and J. Zeng, "Global convergence of admm in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, Jan 2019. [Online]. Available: https://doi.org/10.1007/s10915-018-0757-z

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience, 2006.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[10] F. Iutzeler, P. Bianchi, P. Ciblat, and W. Hachem, "Asynchronous distributed optimization using a randomized alternating direction method of multipliers," *CoRR*, vol. abs/1303.2837, 2013. [Online]. Available: http://arxiv.org/abs/1303.2837

[11] P. Bianchi, W. Hachem, and F. Iutzeler, "A coordinate descent primal-dual algorithm and application to distributed asynchronous optimization," *IEEE Transactions on Automatic Control*, vol. 61, 09 2015.

VI. Appendices

A. Theorem III.1

Proof. To prove the non-convexity of the optimization, we show that the functions $I(X;Y)$ and $I(U;Y)$ are both convex functions of $p(y|x)$. Then it could be argued that the set $I(U;Y) \geq I_{th}$ represents a non-convex set while the objective function is still convex in regards to $p(y|x)$ and thus the entire problem in Eq.(2) would be a non-convex optimization.

Step 1: In this step, we prove that $I(X;Y)$ is a convex function of $p(y|x)$.

Let us assume two probability matrices $\mathbf{p^{(1)}(y|x)}$ and $\mathbf{p^{(2)}(y|x)}$ over the same set of possible x and y's. Then the mutual information between X and Y assuming the prior probability set $\mathbf{p(x)}$ and the conditional probability matrix $\mathbf{p^{(1)}(y|x)}$ could be assumed to be I_1 and the mutual information between X and Y assuming the prior probability set $\mathbf{p(x)}$ and the conditional probability matrix $\mathbf{p^{(2)}(y|x)}$ could be assumed to be I_2. Then all we need to show is that if I represents the mutual information between X and Y assuming the prior probability set $\mathbf{p(x)}$ and the conditional probability matrix $\omega\mathbf{p^{(1)}(y|x)}+(1-\omega)\mathbf{p^{(2)}(y|x)}, 0 \leq \omega \leq 1$, then $I \leq \omega I_1 + (1-\omega)I_2$.

We assume X is drawn with probability vector $\mathbf{p(x)}$. We assume a binary variable S which is equal to 1 with probability ω and 0 with probability $1-\omega$. If $S=1$, we choose Y using $\mathbf{p^{(1)}(y|x)}$ and otherwise we choose Y using $\mathbf{p^{(2)}(y|x)}$. Under such conditions, we will have $I(X;Y) = I$.

On the one hand, $I(SY;X) = I(Y;X) + I(S;X|Y) \geq I(X;Y) = I$. On the other hand, $I(SY;X) = I(X;S) + I(Y;X|S) = 0 + I(Y;X|S) = \omega I(Y;X|S=1) + (1-\omega)I(Y;X|S=0) = \omega I_1 + (1-\omega)I_2$. So, now we can see that $I \leq \omega I_1 + (1-\omega)I_2$. As a result, $I(X;Y)$ is a convex function of $p(y|x)$.

Step 2: Now, we prove that $I(U;Y)$ is a convex function of $p(y|x)$ as well.

Following the same logic as that of step 1, it is easy to see that $I(U;Y)$ is a convex function of $p(y|u)$. Furthermore, we know that $p(y|u) = \sum_x p(xy|u) = \sum_x p(x|u)p(y|x)$. As a result, we can deduce that $I(U;Y)$ is a convex function of $p(y|x)$ as well. □

B. Theorem III.2

Proof. This proof is quite straightforward. If we assume that $0 \leq \beta \leq 1$, it follows that:

$$\beta \leq 1 \rightarrow \beta I(U;Y) \leq I(U;Y)$$
$$\rightarrow I(X;Y) - \beta I(U;Y) \geq I(X;Y) - I(U;Y) \geq 0 \quad (26)$$

where the final inequality on the right hand side of Eq.(26) is derived from the fact that $U \rightarrow X \rightarrow Y$ forms a Markov chain and thus $I(X;Y) \geq I(U;Y)$. Thus, all we need to do is make certain that $I(X;Y) = 0$. As a result as long as $p(y|x)$ follows a deterministic distribution meaning,

$$p(y|x=a) = \begin{cases} 1 & y = b_a \\ 0 & otherwise \end{cases}, \forall a = 1, ..., |x| \quad (27)$$

the objective function in Eq.(3) can be minimized to 0. In the above equation, a represents any value that the input variable $\mathbf{X}$ could hold and b_a represents the value of output $\mathbf{Y}$ which will be deterministically chosen once the input takes the a value. □